7 Paskaita. Neparametriniai kriterijai. Pirsono (Chi kvadratu) kriterijus.

7.1 Chi kvadratu suderinamumo kriterijus.

Statistinės hipotezės, apibūdinančios tiriamo atsitiktinio dydžio skirstinį, o ne konkrečias parametrų reikšmes, vadinamos neparametrinėmis. Pvz., ar tiriamo atsitiktinio dydžio skirstinys yra normalusis, Puasono, binominis, t.t. Nagrinėdami neparametrines hipotezes, reikia išskirti 2 atvejus: kai stebimo atsitiktinio dydžio tikėtinas skirstinys nepriklauso nuo nežinomų parametrų ir kai jis priklauso nuo vieno ar kelių nežinomų parametrų.

1. **Pirmasis modelis.** Tarkime, kad atlikus eksperimentą, gali įvykti vienas iš k nesutaikomų įvykių $A_1, A_2, \ldots, A_k$ su atitinkamomis tikimybėmis $p_1, p_2, \ldots, p_k, p_1 + p_2 + \ldots + p_k = 1$. Atliekame n nepriklausomų eksperimentų. Pažymėkime $O_i, i = 1, 2, \ldots, k$ (nuo angl. *observed*), stebimus įvykio A_i pasirodymų skaičius, atlikus n eksperimentų. Tikėtini įvykių A_i dažniai, atlikus n eksperimentų, skaičiuojami pagal formulę: $E_i = np_i$ (nuo angl. *expected*, tikėtinas). Statistika

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(O_i - np_i)^2}{np_i}, \tag{1}$$

   kai n neaprėžtai didėja, turi χ^2 skirstinį su $k-1$ laisvės laipsniais.

2. **Antrasis modelis.** Šis modelis nuo pirmojo modelio skiriasi tuo, kad tikimybės p_i priklauso nuo s nežinomų parametrų $p_i = p_i(\alpha_1, \ldots, \alpha_s)$. Apskaičiavę nežinomų parametrų taškinius įverčius $\hat{\alpha_1}, \ldots, \hat{\alpha_s}$, gauname tokia statistikos X^2 išraišką:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - np_i(\hat{\alpha_1}, \ldots, \hat{\alpha_s}))^2}{np_i(\hat{\alpha_1}, \ldots, \hat{\alpha_s})}. \tag{2}$$

   Kai n neaprėžtai didėja, X^2 turi χ^2 skirstinį su $k-s-1$ laisvės laipsniais.

Kai atsitiktinis dydis yra pasiskirstęs pagal hipotetinį teorinį modelį, skirtumai tarp stebimų ir tikėtinų dažnių yra nedideli, todėl statistikos X^2 reikšmė yra nedidelė. Todėl šiuo atveju tikėtina, kad teorinis modelis yra tinkamas apibūdinant stebimą atsitiktinį dydį. Hipotezei apie stebimo atsitiktinio dydžio skirstinį tikrinti naudojamas kriterijus, kuris dar vadinamas *Pirsono suderinamumo kriterijumi*. χ^2 kriterijus parodo, ar skirtumas tarp empirinio ir teorinio skirstinių yra reikšmingas, t.y. ar stebimas empirinis skirstinys suderinamas su teoriniu modeliu.

Diskretusis atvejis. Tarkime, kad atlikus n eksperimentų, stebimas diskretusis atsitiktinis dydis įgyja k skirtingų reikšmių x_i su atitinkamais dažniais

$O_i, i = 1, 2, \ldots, k, O_1 + \ldots + O_k = n$. Tikrinsime hipotezę, kad stebimo kintamojo skirstinys yra žinomas ir tikimybė, kad jis įgis reikšmę x_i lygi p_i^0. Tikėtini dažniai, atlikus n eksperimentų lygūs $E_i = np_i^0$. Tiek stebėjimų turėtų priklausyti i-jai kategorijai. Iš tikrųjų šiai kategorijai priklauso O_i imties stebėjimų. Tikrinama hipotezė:

$$\begin{cases} H_0 : p_1 = p_1^0, \ldots, p_k = p_k^0, \\ H_1 : p_i \neq p_i^0 \text{ bent vienam } i. \end{cases}$$

Pirsono suderinamumo kriterijaus statistika (1), kai nulinė hipotezė yra teisinga, turi χ^2 skirstinį su $k-1$ laisvės laipsniais. Tarkime, reikšmingumo lygmuo lygus α. Tuomet hipotezė H_0 atmetama (t.y. darome išvadą, kad prielaida apie kintamojo teorinį skirstinį populiacijoje nepasitvirtino), jei $X^2 > X_\alpha^2(k-1)$, čia $X_\alpha^2(k-1)$ yra χ^2 skirstinio su $k-1$ laisvės laipsniais α lygmens kritinė reikšmė.

Pavyzdys. Patikrinkite hipotezę ($\alpha = 0.05$), kad tarp visų butų, parduodamų Vilniaus mieste:
vieno kambario butai sudaro 1/7 rinkos dalį;
dviejų kambarių butai sudaro 3/7 rinkos dalį;
trijų kambarių butai sudaro 2/7 rinkos dalį;
keturių kambarių butai sudaro 1/7 rinkos dalį;
Tarp laikraštyje esančių 392 skelbimų apie buto pardavimą buvo rasti tokie skelbimai:

1 lentelė: Parduodami butai Vilniaus mieste

x_i	1 kambario	2 kambarių	3 kambarių	4 kambarių
O_i	45	176	100	71

Sprendimas. Tikriname hipotezę:

$$\begin{cases} H_0 : p_1 = p_4 = 1/7, p_2 = 3/7, p_3 = 2/7, \\ H_1 : p_i \neq p_i^0 \text{ bent vienam } i. \end{cases}$$

Randame tikėtinus dažnius pagal formulę $E_i = np_i^0$:

$E_1 = 392 \cdot 1/7 = 56, E_2 = 392 \cdot 3/7 = 168, E_3 = 392 \cdot 2/7 = 112, E_4 = 392 \cdot 1/7 = 56.$

Kriterijaus statistika:

$$X^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{(45-56)^2}{56} + \frac{(176-168)^2}{168} + \frac{(100-112)^2}{112} + \frac{(71-56)^2}{56} = 7,845.$$

$X_{0,05}^2(3) = 7,815$. Kadangi $7,845 > X_\alpha^2(k-1)$, H_0 atmetame. Proporcijos yra kitokios.

Tolydusis atvejis. Normalusis skirstinys. Tikrinama hipotezė apie tai, kad stebimas tolydusis atsitiktinis dydis turi normalųjį skirstinį:

$$\begin{cases} H_0 : X \sim \mathcal{N}(\mu, \sigma^2), \\ H_1 : X \text{ skirstinys yra kitoks.} \end{cases}$$

Stebimas kintamasis yra tolydusis, todėl norėdami taikyti χ^2 kriterijų, turime jį sudiskretinti. Pradiniai duomenys grupuojami į k intervalų. Tuomet kiekviena stebima reikšmė gali patekti tik į vieną intervalą, t.y. gali įvykti vienas iš k nesutaikomų įvykių $A_1, \ldots, A_k$. Čia A_i – X reikšmė pateko į i-jį intervalą. Turime apskaičiuoti tikimybes, kad normaliojo atsitiktinio dydžio reikšmė pateks į i-jį intervalą p_i ir tikėtinus dažnius $E_i = np_i$. Turime antrąjį modelį, kai teorinis skirstinys priklauso nuo dviejų nežinomų parametrų μ, σ^2. Hipotezei tikrinti naudojama statistika (2), kuri turi χ^2 skirstinį su $k-3$ laisvės laipsniais, jei duomenys gauti stebint normalųjį atsitiktinį dydį, kai n neaprėžtai didėja.

Sakykime, galimų reikšmių aibė suskirstyta į k nesikertančių intervalų $(-\infty; c_1), [c_1, c_2), \ldots, [c_{k-1}; \infty)$ ir į i-jį intervalą pateko o_i stebėjimų. Kriterijaus statistikos realizacija:

$$x^2 = \sum_{i=1}^{k} \frac{(o_i - np_i(\bar{x}, s_1))^2}{np_i(\bar{x}, s_1)},$$

čia $p_i = \Phi\left(\frac{c_i - \bar{x}}{s_1}\right) - \Phi\left(\frac{c_{i-1} - \bar{x}}{s_1}\right)$. Tarkime, reikšmingumo lygmuo lygus α. Tuomet hipotezė H_0 atmetama (t.y. duomenys neleidžia teigti, kad stebimas atsitiktinis dydis turi normalųjį skirstinį), jei $X^2 > X^2_\alpha(k-3)$, čia $X^2_\alpha(k-3)$ yra χ^2 skirstinio su $k-3$ laisvės laipsniais α lygmens kritinė reikšmė.

Pavyzdys. Duota tolydžiojo atsitiktinio dydžio grupuotų duomenų imtis: Pasinaudodami Pirsono suderinamumo kriterijumi, patikrinkite, ar imtis yra iš

2 lentelė: Stebimi duomenys

c_{i-1}	c_i	x_i^*	o_i
3	8	5.5	6
8	13	10.5	8
13	18	15.5	15
18	23	20.5	40
23	28	25.5	16
28	33	30.5	8
33	38	35.5	7

populiacijos, turinčios normalųjį skirstinį ($\alpha = 0,05$).

Sprendimas. Imties elementų skaičius $n = 100$. Randame imties vidurkį, dispersiją ir standartinį nuokrypį:

$$\bar{x} = \frac{1}{100}(5.5 \cdot 6 + 10.5 \cdot 8 + \ldots 35.5 \cdot 7) = 20,7, \quad s_1^2 = 51,41, s_1 = 7,17.$$

Surašykime tarpinius rezultatus į pagalbinę lentelę: Čia pasinaudojome tuo, kad $\Phi(-x) = -\Phi(x)$. $x^2 = 13,698, X^2_{0,05}(7-3) = X^2_{0,05}(4) = 9,5$. Kadangi

3 lentelė: Pagalbinė lentelė.

$\frac{c_{i-1}-\bar{x}}{s_1}$	$\frac{c_i-\bar{x}}{s_1}$	$\Phi\left(\frac{c_{i-1}-\bar{x}}{s_1}\right)$	$\Phi\left(\frac{c_{i-1}-\bar{x}}{s_1}\right)$	o_i	$e_i = np_i$	$o_i - e_i$	$\frac{(o_i-e_i)^2}{e_i}$
$-\infty$	-1.771	0	0.038	6	3.826	2.174202	1.2356
-1.771	-1.074	0.038	0.141	8	10.32	-2.31715	0.520413
-1.074	-0.377	0.141	0.353	15	21.18	-6.18175	1.804101
-0.377	0.321	0.353	0.626	40	27.26	12.74351	5.958108
0.321	1.018	0.626	0.846	16	21.99	-5.98803	1.63073
1.018	1.715	0.846	0.957	8	11.12	-3.11793	0.874398
1.715	∞	0.957	1	7	4.313	2.687148	1.674243
Iš viso:							13.698

$13,698 > X^2_{0,05}(4)$, hipotezė apie tai, kad duomenys gauti iš normaliosios populiacijos, atmetama. Chi kvadratu suderinamumo kriterijaus taikymo pavyzdys R pateiktas pav. 1.

```
> x <- c(89,37,30,28,2)
> p <- c(40,20,20,15,5)
> chisq.test(x, p = p)
Error in chisq.test(x, p = p) : probabilities must sum to 1.
> chisq.test(x, p = p, rescale.p = TRUE)

        Chi-squared test for given probabilities

data:  x
X-squared = 9.9901, df = 4, p-value = 0.04059

> p <- c(0.40,0.20,0.20,0.15,0.05)
> chisq.test(x, p = p)

        Chi-squared test for given probabilities

data:  x
X-squared = 9.9901, df = 4, p-value = 0.04059

> p <- c(0.40,0.20,0.20,0.19,0.01)
> chisq.test(x, p = p)

        Chi-squared test for given probabilities

data:  x
X-squared = 5.7947, df = 4, p-value = 0.215

Warning message:
In chisq.test(x, p = p) : Chi-squared approximation may be incorrect
```

1 pav.: Chi kvadratu suderinamumo kriterijaus taikymas R

Chi kvadratu kriterijaus privalumai – jis gali būti taikomas, kai stebimo kintamojo skirstinys nėra normalusis (šis reikalavimas yra būtinas, kai taikome parametrinius kriterijus). Kitas privalumas – kriterijus tinka ir diskretiesiems ir tolydiesiems skirstiniams, be to, stebėjimų skaičius neturi būti labai didelis. Yra tam tikrų reikalavimų kriterijaus taikymui. Stebėjimų skaičius n turi būti

nemažesnis už 30; bent 80% dažnių lentelės tikėtinų dažnių turi būti ne mažesni kaip 5; neturi būti langelių su nuliniais tikėtinais dažniais. Jei dvi paskutinės sąlygos netenkinamos, galima mažinti kategorijų skaičių (jungti kategorijas).

7.2 Chi kvadratu nepriklausomumo kriterijus.

Tarkime, kad stebime diskrečių atsitiktinių dydžių porą (X, Y). Atsitiktinis dydis X įgyja I skirtingų reikšmių, Atsitiktinis dydis Y įgyja J skirtingų reikšmių. 2 pav. pavaizduota dviejų požymių priklausomumo lentelė. Čia o_{ij} – stebimi dažniai. Norime patikrinti hipotezę, kad kintamieji X ir Y yra nepriklausomi.

X\Y	y_1	y_2	...	y_J	Σ
x_1	o_{11}	o_{12}	...	o_{1J}	$o_{1\bullet}$
x_2	o_{21}	o_{22}	...	o_{2J}	$o_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_I	o_{I1}	o_{I2}	...	o_{IJ}	$o_{I\bullet}$
Σ	$o_{\bullet 1}$	$o_{\bullet 2}$	...	$o_{\bullet J}$	n

2 pav.: Dviejų požymių priklausomumo lentelė.

Pažymėkime

$$p_{ij} = P(X = x_i, Y = y_j), p_i = P(X = x_i), q_j = P(Y = y_j).$$

Iš tikimybių teorijos žinome, kad kintamieji yra nepriklausomi, jei

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$$

su visais $i = 1, 2, \ldots, I, j = 1, 2, \ldots, J$. Taigi mums reikia patikrinti hipotezę

$$\begin{cases} H_0 : p_{ij} = p_i q_j, \forall i, j, \\ H_1 : p_{ij} \neq p_i q_j \text{ bent vienai porai } (i, j). \end{cases}$$

Naudosime tą pačią Chi kvadratu kriterijaus statistiką (1), kur tikėtini dažniai apskaičiuojami taip:

$$E_{ij} = n\hat{p_{ij}} = n\hat{p}_i\hat{q}_j = n\frac{o_{i\cdot}}{n}\frac{o_{\cdot j}}{n} = \frac{o_{i\cdot}o_{\cdot j}}{n}.$$

Statistika (1), turi asimptotiškai χ^2 skirstinį su $(I-1)(J-1)$ laisvės laipsniais, kai H_0 teisinga.

Pavyzdys. Buvo apklausti 400 studentų. Kiekvienas studentas vertino tik po vieną dėstytoją. Ar studentai dėstytojus vertina priklausomai nuo gautų pažymių? $\alpha = 0,05$?

	Studento gautas balas				
Vertinimas	**0-4**	**5-6**	**7-8**	**9**	**10**
netikęs	40	45	10	10	15
vidutiniškas	30	40	60	25	25
puikus	10	15	30	25	20

3 pav.: Stebimų dažnių lentelė.

	Studento gautas balas					
Vertinimas	**0-4**	**5-6**	**7-8**	**9**	**10**	
netikęs	24	30	30	18	18	**120**
vidutiniškas	36	45	45	27	27	**180**
puikus	20	25	25	15	15	**100**
	80	**100**	**100**	**60**	**60**	**400**

4 pav.: Tikėtinų dažnių lentelė.

Sprendimas. Pirmiausia apskaičiuojami tikėtini dažniai tokiu būdu:

$$E_{11} = \frac{80 \cdot 120}{400} = 24, \ldots, E_{34} = \frac{60 \cdot 100}{400} = 15.$$

Apskaičiuojama kriterijaus statistikos reikšmė. Kriterijaus statistika turi χ^2 skirstinį su $(5-1)(3-1) = 8$ laisvės laipsniais, kai H_0 teisinga.

$$X^2 = \frac{(40-24)^2}{24} + \ldots + \frac{(25-15)^2}{15} = 60,741, X^2_{0,05}(8) = 15,51.$$

Kadangi $60,741 > X^2_{0,05}(8)$, hipotezė apie tai, kad požymiai X ir Y yra nepriklausomi, atmetama. Išvada: studento gautas pažymys turi įtakos dėstytojo vertinimui.

7.3 Chi kvadratu homogeniškumo kriterijus.

Tarkime, kad keliose populiacijose stebimas vienas ir tas pats diskretusis kintamasis X. Reikia patikrinti hipotezę, kad kintamojo X skirstinys visose populiacijose vienodas. Šis kriterijus naudojamas, kai, pavyzdžiui, reikia atsakyti į klausimus:

1. Ar rūkančių vyrų ir moterų procentas yra tas pats?

2. Ar požiūris į lengvų narkotikų legalizavimą įvairiose amžiaus grupėse yra toks pat?

3. Ar Lietuvoje gyvenančių įvairių tautybių žmonių išsilavinimas yra vienodas?

Šios hipotezės tikrinimas atliekamas taip pat, kaip ir nepriklausomumo hipotezės tikrinimas.

Pavyzdys. Tiriama, ar vyrai ir moterys vienodai vertina, jei jų darbdavys priimtų į darbą grįžusius iš įkalinimo vietų. Buvo apklausta 200 atsitiktinai parinktų vyrų ir 100 atsitiktinai parinktų moterų. Atsakymų variantai: teigiamai, neturiu nuomonės, neigiamai. $\alpha = 0,05$.

	teigiamai	neturiu nuomonės	neigiamai	Iš viso:
vyrai	60 (56)	70 (67,3)	70 (76,7)	200
moterys	24 (28)	31 (33,7)	45 (38,3)	100
Iš viso:	84	101	115	300

5 pav.: Stebimų ir tikėtinų (skliausteliuose) dažnių lentelė.

Sprendimas. Tikrinama hipotezė:

$$\begin{cases} H_0 : p_{11} = p_{21}, p_{12} = p_{22}, p_{13} = p_{23}, \\ H_1 : p_{1j} \neq p_{2j} \text{ bent vienam } j. \end{cases}$$

Randame tikėtinus dažnius pagal formulę $E_{ij} = \frac{o_{i\cdot} o_{\cdot j}}{n}$, pavyzdžiui,

$$E_{11} = \frac{84 \cdot 200}{300} = 56.$$

Apskaičiuojama kriterijaus statistikos reikšmė. Kriterijaus statistika turi χ^2 skirstinį su $(2-1)(3-1) = 2$ laisvės laipsniais, kai H_0 teisinga.

$$X^2 = 2,939, X^2_{0,05}(2) = 5,99.$$

Kadangi $2,93 < X^2_{0,05}(2)$, H_0 neatmetama. Išvada: vyrų ir moterų požiūris į grįžusių iš įkalinimo vietų įdarbinimą jų įmonėje yra vienodas.